

ANÁLISIS DE REGRESIÓN

Cap.4.4: Regresión Lineal Múltiple (RLM)

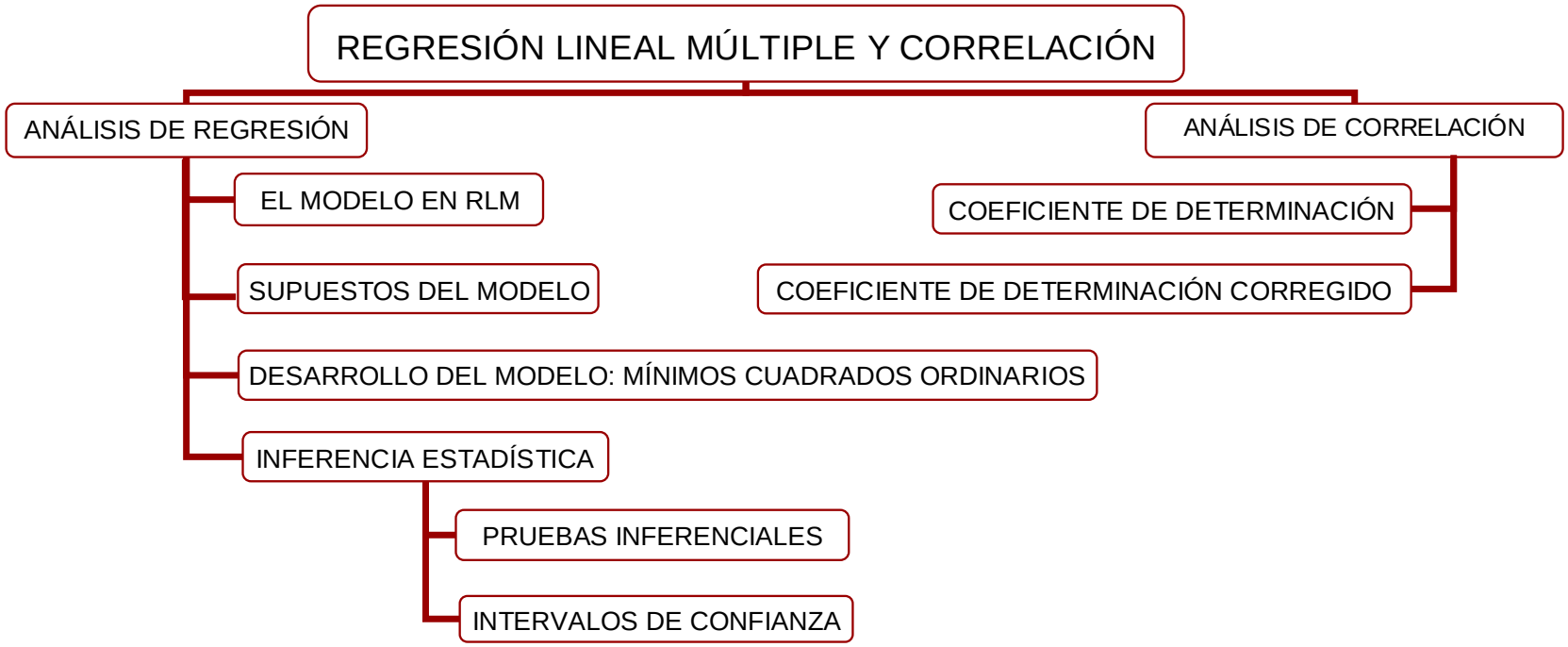
① El modelo

② Mínimos cuadrados

③ Inferencia estadística

INTRODUCCIÓN

En un modelo de regresión, al utilizar más de una variable independiente es posible incrementar el poder explicativo y la utilidad del modelo en la toma de muchas decisiones de negocios. Esta sección analiza dichos «Modelos de Regresión Lineal Múltiple».



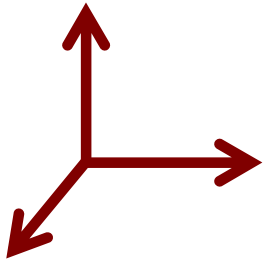
REGRESIÓN LINEAL MULTIPLE

El procedimiento de Regresión Lineal permite utilizar más de una variable independiente y permite llevar a cabo análisis de regresión múltiple.

En el análisis de regresión múltiple la ecuación ya no define una recta en el plano, sino un *hiperplano en un espacio multidimensional.*

Con una variable dependiente y dos independientes…

…necesitamos tres ejes para poder representar el diagrama de dispersión



Con más de una variable independiente, la representación gráfica de las relaciones presentes en un modelo de regresión resulta *poco intuitiva, muy complicada y nada útil*

Sistema coordenado rectangular en tres dimensiones

El sistema coordenado rectangular en tres dimensiones o R^3 se forma cuando tres ejes de números reales mutuamente perpendiculares se intersecan en el origen de cada eje.

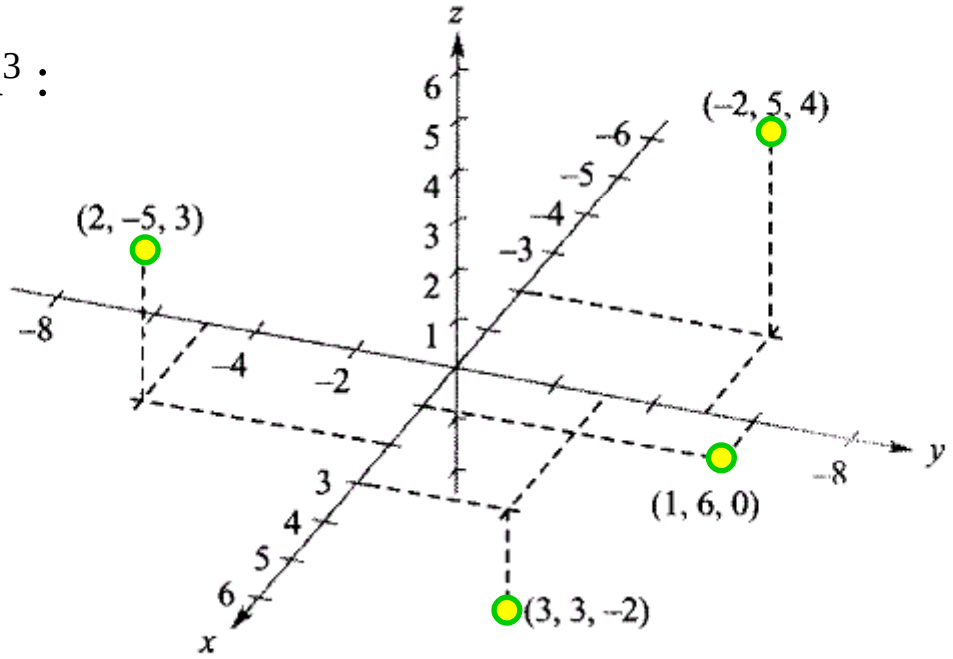
Ubicación de un punto en R^3 :

(1,6,0)

(3,3,-2)

(-2,5,4)

(2,-5,4)



En un sistema de coordenadas tridimensional los puntos se representan por tríos ordenados.

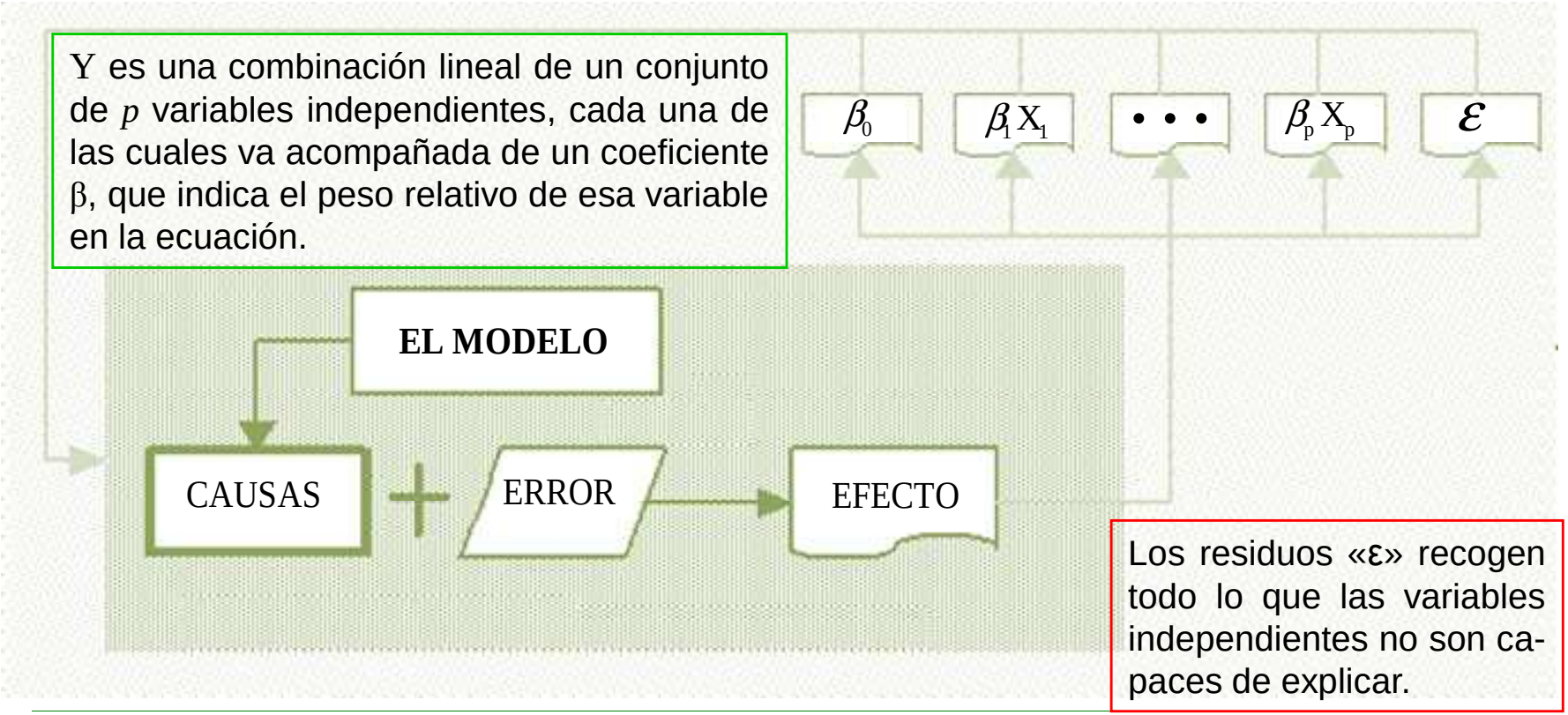
Fuente: Larson Vol 2

EL MODELO DE REGRESIÓN LINEAL MÚLTIPLE

Es más fácil y práctico partir de la ecuación del MRLM:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

Y es una combinación lineal de un conjunto de *p* variables independientes, cada una de las cuales va acompañada de un coeficiente β, que indica el peso relativo de esa variable en la ecuación.



Los residuos «ε» recogen todo lo que las variables independientes no son capaces de explicar.

EL MODELO DE REGRESIÓN LINEAL MÚLTIPLE

El modelo de regresión lineal múltiple con *p* variables predictoras y basado en *n* observaciones está dado por:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \mathrm{e}_i \text{ para i = 1,2,…,}n$$

$$\begin{pmatrix} \mathrm{Y}_1 \\ \mathrm{Y}_2 \\ \vdots \\ \mathrm{Y}_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \mathrm{e}_1 \\ \mathrm{e}_2 \\ \vdots \\ \mathrm{e}_n \end{pmatrix}$$

← Forma matricial

$$Y_{n\times 1} = \mathrm{X}_{\mathrm{n}\times(\mathrm{p}+1)} \beta_{(\mathrm{p}+1)\times 1} + \varepsilon_{\mathrm{n}\times 1}$$

(Parte determinística) (Parte Aleatoria)

SUPUESTOS DEL MODELO DE REGRESIÓN LINEAL MÚLTIPLE

Para el modelo de regresión lineal múltiple con *p* variables predictoras y basado en *n* observaciones, deben considerarse los siguientes supuestos cuando se realiza el análisis de regresión:

① Las variables X's toman valores distintos en la muestra.

② Los residuos, e_i, tienen media cero y varianza constante; este último supuesto es conocido como «HOMOCEDASTICIDAD».

③ Los residuos, e_i, son estadísticamente independientes entre sí.

④ La distribución de los residuos, e_i, es normal.

Estimación del modelo de regresión lineal múltiple

Para las n observaciones de la muestra, la ecuación de regresión lineal múltiple es:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip} \quad \text{para} \quad i = 1,2,\ldots,n$$

En términos de matrices, estas n ecuaciones se describen como

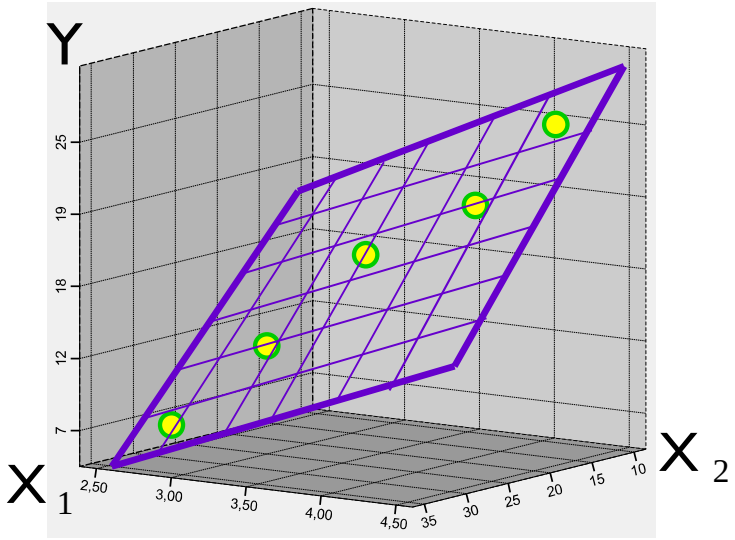
$$\hat{\mathrm{Y}}_{\mathrm{n}\times 1} = X_{n\times(p+1)}\hat{\beta}_{(p+1)\times 1}$$

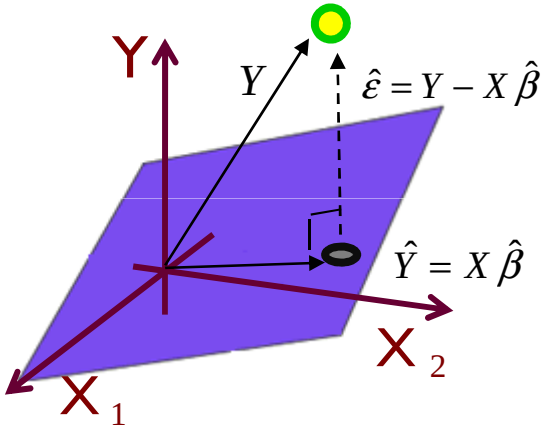
Matriz de datos de las variables independientes

Vector de estimadores del vector β

Diagrama de dispersión: Ajuste con matrices

Si representamos los puntos (datos de la muestra) en un diagrama de dispersión o “nube de puntos” tendremos una idea gráfica de la posible relación entre Y y el conjunto de variables predictoras.





En conclusión, por lo expuesto, uno de los objetivos en el *Análisis de Regresión Múltiple* es hallar entre todos los posibles vectores “beta techo” los que minimicen la suma de los residuos(errores) al cuadrado.

Estimación del vector β por mínimos cuadrados

Consiste en minimizar la suma de cuadrados de los errores (SCE):

$$\text{SCE} = Q\left(\hat{\beta}\right) = \text{e}' \cdot \text{e} = \left(\text{Y-X}\hat{\beta}\right)'\left(\text{Y} - \text{X}\hat{\beta}\right)$$

Haciendo operaciones con los vectores y matrices

$$Q(\boldsymbol{\beta}) = \mathbf{Y'Y} - \boldsymbol{\beta}'\mathbf{X'Y} - \mathbf{Y'X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X'X}\boldsymbol{\beta} = \mathbf{Y'Y} - 2\boldsymbol{\beta}'\mathbf{X'Y} + \boldsymbol{\beta}'\mathbf{X'X}\boldsymbol{\beta}$$

Derivando Q con respecto a $\hat{\beta}$ e igualando a cero se obtiene el sistema de ecuaciones normales:

$$\text{X'X}\hat{\beta} = \text{X'Y}$$

Estimación del vector β por mínimos cuadrados

donde,

$$X'X = \begin{pmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \cdots & \sum_{i=1}^{n} x_{ip} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & \cdots & \sum_{i=1}^{n} x_{i1}x_{ip} \\ \vdots & \vdots & & & \vdots \\ \sum_{i=1}^{n} x_{ip} & \sum_{i=1}^{n} x_{ip}x_{i1} & \sum_{i=1}^{n} x_{ip}x_{i2} & \cdots & \sum_{i=1}^{n} x_{ip}^2 \end{pmatrix}; \quad X'Y = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ip}y_i \end{pmatrix}$$

Si la matriz $X'X$ es no singular, resolviendo para $\hat{\beta}$ se obtiene:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Estimación del vector β por mínimos cuadrados

ERROR ESTANDAR DE LA ESTIMACIÓN

El error estándar de la estimación mide la variabilidad de los valores muestrales y observados alrededor de la ecuación de regresión.

ERROR ESTÁNDAR DE LA ESTIMACIÓN →

$$s = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - p - 1}} = \sqrt{\frac{\text{SCE}}{n - p - 1}} = \sqrt{\text{CME}}$$

También llamado error típico de la estimación.

Propiedades de los estimadores de mínimos cuadrados

1. Las estadísticas $\hat{\beta}_i$ son estimadores insesgados de β_i.

2. La matriz $\sigma^2(\mathrm{X'X})^{-1}$ se denomina matriz de VARIANZAS - COVARIANZAS de vector $\hat{\beta}$.

   Por ejemplo, para $p=2$ variables independientes, se tiene:

$$\sigma^2(\mathrm{X'X})^{-1} = \sigma^2 \begin{pmatrix} c_{00} & c_{01} & c_{02} \\ c_{10} & c_{11} & c_{12} \\ c_{20} & c_{21} & c_{22} \end{pmatrix} \xrightarrow[i=0,1,2]{\text{para...}} \begin{array}{l} \mathrm{Var}(\hat{\beta}_i) \quad \text{o} \quad \sigma^2_{\hat{\beta}_i} = c_{ii}\sigma^2 \quad \text{y} \\ \mathrm{Cov}(\hat{\beta}_i, \hat{\beta}_j) = c_{ij}\sigma^2 \ \text{ para } \ i \neq j \end{array}$$

- $\mathrm{Var}(\hat{\beta}_i) \equiv$ varianza del estimador del coeficiente de regresión parcial β_i

- $\mathrm{Cov}(\hat{\beta}_i, \hat{\beta}_j) \equiv$ covarianza entre los estimadores de los coeficientes de regresión parcial β_i y β_j

Propiedades de los estimadores de mínimos cuadrados

3. Una estimación insesgada de la varianza σ^2 es la varianza muestral (s^2 o $\hat{\sigma}^2$) .

   El error estándar de estimación múltiple es $s = \hat{\sigma} = \sqrt{s^2}$.

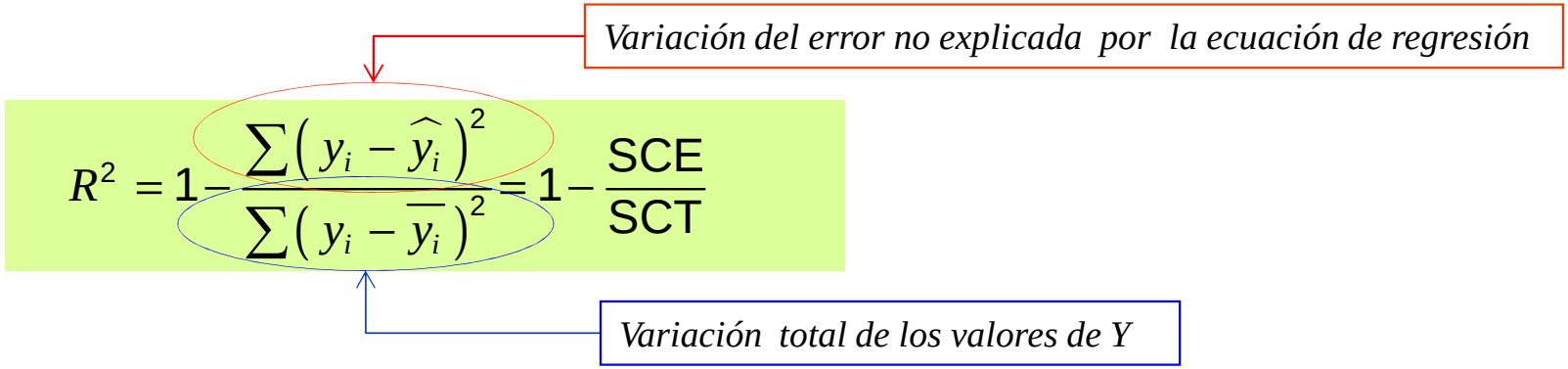
4. La estimación de la varianza $\text{Var}\left(\hat{\beta}_i\right)$ o $\sigma^2_{\hat{\beta}_i} = c_{ii}\sigma^2$ es entonces $\hat{\sigma}^2_{\hat{\beta}_i} = c_{ii}\hat{\sigma}^2$.

   El error estándar de cada $\hat{\beta}_i$ es

$$s.e.\left(\hat{\beta}_i\right) = \hat{\sigma}_{\hat{\beta}_i} = \sqrt{c_{ii}\hat{\sigma}^2} = s\sqrt{c_{ii}}$$

EL COEFICIENTE DE DETERMINACIÓN MÚLTIPLE

El coeficiente de determinación múltiple, R^2, es una medida de la bondad de ajuste del modelo de regresión y se determina por:



- R^2 varía entre 0 y 1, y es bastante común expresarlo en porcentaje.
- Un R^2 mayor del 70 % indica una buena asociación lineal entre las variables, luego las variables X's puede usarse para predecir Y.

Recordar que «R^2» indica qué porcentaje de la variabilidad de la variable de respuesta Y es explicada por el modelo de regresión.

Descomposición de la variación total de Y

La variación total de Y es la suma de la variación explicada más la variación no explicada (el error):

$$SCT = SCR + SCE$$

EL COEFICIENTE DE DETERMINACIÓN MÚLTIPLE AJUSTADO

El coeficiente de determinación corregido por grados de libertad o coeficiente de determinación ajustado $\hat{R}^2$, está dado por:

$$\hat{R}^2 = 1 - \frac{SCE/(n-p-1)}{SCT/(n-1)} = 1 - \frac{CME}{CMT}$$

Prueba de ANOVA para la significancia general de la ecuación de regresión múltiple

1. Planteamiento:

la ecuación de regresión no es significativa

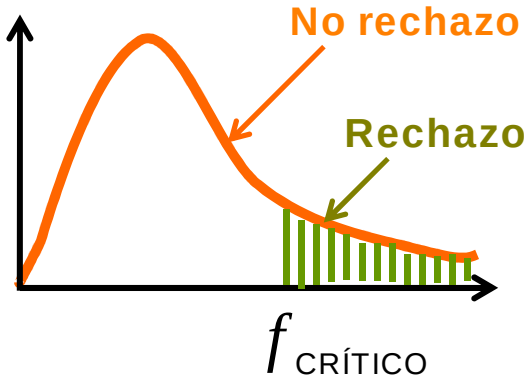
$H_o: \beta_1=\beta_2=\cdots=\beta_p=0$

la ecuación de regresión es significativa

$H_1:$ uno o más de los valores de $\beta_i \neq 0 \ (i=1,\ldots,p)$

2. Estadístico de prueba: $F = \dfrac{SCR/p}{SCE/(n-p-1)}$

3. Región crítica. Decisión:



$f_{CRÍTICO} = F_{\alpha;\ p\ \wedge\ n-p-1\ \text{g.l.}}$ (Tabla)

$f_{CRÍTICO} = F_{\alpha;\ p\ \wedge\ n-p-1\ \text{g.l.}}$ (Excel)

TABLA DE ANÁLISIS DE VARIANZA

En esta tabla se descompone la variabilidad de la respuesta en función de la variabilidad explicada y no explicada por la regresión ajustada. También se obtiene el valor del estadístico de contraste F.

ANÁLISIS DE VARIANZA

Origen de las variaciones	*Suma de cuadrados*	*Grados de libertad*	*Promedio de los cuadrados*	*F*
Debido a la regresión	SCR	p	CMR=SCR/p	CMR/CME
Debido al error	SCE	n-p-1	CME=SCE/(n-p-1)	
Total	SCT	n-1		

Prueba *t* de significancia de un coeficiente de regresión parcial $\hat{\beta}_i$

1. Planteamiento:

H_0: $\beta_i = 0$ (el coeficiente poblacional es cero)

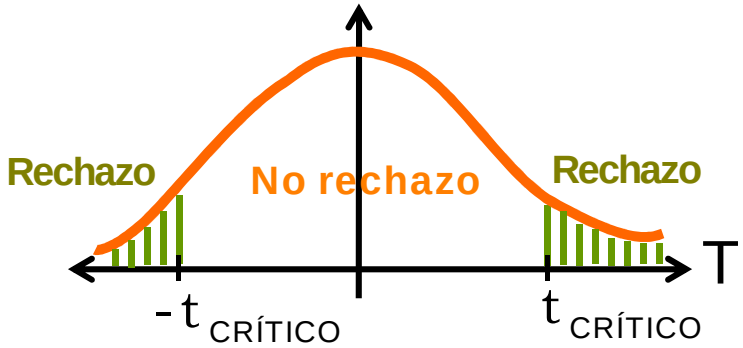
H_1: $\beta_i \neq 0$ (el coeficiente poblacional es diferente de cero)

2. Estadístico:

$$T = \frac{\hat{\beta}_i}{s.e.\left(\hat{\beta}_i\right)} \sim t_{(n-p-1\ \text{g.l.})}$$

DESVIACIÓN ESTÁNDAR ESTIMADA DE $\hat{\beta}_i$ (ERROR TÍPICO)

3. Región crítica. Decisión:



$$t_{CRÍTICO} = \pm t_{1-\alpha/2;\ n-p-1\ \text{g.l.}} \quad (\text{Tabla})$$

$$t_{CRÍTICO} = \pm t_{\alpha;\ n-p-1\ \text{g.l.}} \quad (\text{Excel})$$

Estimación de intervalo para los coeficiente de regresión parcial

El intervalo de confianza del 100(1-α)% para un coeficiente de regresión parcial β_i es

$$I.C.(\beta_i) = \hat{\beta}_i \pm t_{(1-\alpha/2;\, n-p-1)}\, s.e.(\hat{\beta}_i)$$

Desviación estándar estimada de $\hat{\beta}_i$ (error típico)

Nota:

La decisión de rechazar o no rechazar $\beta_i = 0$ (hipótesis nula) se puede realizar también a partir del intervalo de confianza para β_i , verificando si dicho intervalo contiene o no el valor 0.

Intervalos de confianza y de predicción en RLM

Se desea predecir el valor medio de la variable de respuesta Y para una combinación predeterminada de las variables predictoras $X_1,\ldots,X_p$.

Consideremos el vector de valores observados $\mathbf{x}'_0 = (1, x_{1,0},\ldots,x_{p,0})$

El valor predicho para el valor medio de la variable de respuesta Y será $\hat{y}_o = \mathbf{x}'_\mathbf{o}\,\hat{\boldsymbol{\beta}}$ y $\mathrm{Var}(\hat{y}_o) = \mathbf{x}'_\mathbf{o}\,\mathbf{Var}(\hat{\boldsymbol{\beta}})\mathbf{x}_\mathbf{o} = \sigma^2\mathbf{x}'_\mathbf{o}\,(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_\mathbf{o}$.

Se asume que los errores están normalmente distribuidos.

Intervalos de confianza y de predicción en RLM

Un intervalo del 100(1-α)% para el valor medio de Y dado que $x = \mathbf{x}'_0$ es de la forma

$$\hat{y}_o \pm t_{(1-\alpha/2,\,n-p-1)} \cdot s \cdot \sqrt{x'_o (X'X)^{-1} x_o}$$

Un intervalo de confianza (intervalo de predicción) del 100(1-α)% para el valor individual de Y dado $x = \mathbf{x}'_0$ es:

$$\hat{y}_o \pm t_{(1-\alpha/2,\,n-p-1)} \cdot s \cdot \sqrt{1 + x'_o (X'X)^{-1} x_o}$$

Ejemplo: predicción del número de turistas

A partir de la siguiente información de una localidad turística sobre los turistas extranjeros llegados de 5 países de procedencia:

PAÍS	Y (dependiente)	X_1 (independiente)	X_2 (independiente)
	Número de turistas	Ingresos medios anuales (miles de euros)	Distancia (cientos de km)
A	18	5	17
B	25	10	15
C	7	2	32
D	12	4	25
E	19	6	20



a) Determine las ecuaciones normales $X'X\hat{\beta} = X'Y$.
b) Determine el vector solución de las ecuaciones normales $\hat{\beta}$ y estime el modelo de regresión lineal múltiple.
c) Calcule s^2 la estimación insesgada de la varianza de la regresión múltiple poblacional.

Ejemplo (continua)

d) Halle la varianza del vector $\hat{\beta}$, $\text{Var}\left(\hat{\beta}\right)$, y el error estándar de cada uno de los coeficientes de la regresión muestral.

e) Determine el coeficiente de determinación múltiple y el de determinación múltiple ajustado. Interprete sus resultados.

f) Al nivel de significación del 10%, realice una prueba de hipótesis general para determinar si alguno de los coeficientes de regresión poblacional no es igual a cero.

g) Al nivel de significación del 10%, realice una prueba de hipótesis para los coeficientes individuales. ¿Qué variable eliminaría?

h) Construya el intervalo de confianza de 90% para cada coeficiente individual.

i) Determine el intervalo de confianza de 90% para estimar el número promedio y el valor particular de Y si $X_1 = 8,\ X_2 = 16$.